AEM
Journals.ASM.org

Genome Sequence Analysis Indicates that the Model Eukaryote *Nematostella vectensis* Harbors Bacterial Consorts

Irena I. Artamonova,[a,b,c] Arcady R. Mushegian[d,e]*

N. I. Vavilov Institute of General Genetics RAS, Moscow, Russia[a]; A. A. Kharkevich Institute for Information Transmission Problems RAS, Moscow, Russia[b]; M. V. Lomonosov Moscow State University, Faculty of Bioengineering and Bioinformatics, Moscow, Russia[c]; Stowers Institute for Medical Research, Kansas City, Missouri, USA[d]; Department of Microbiology, Kansas University Medical Center, Kansas City, Kansas, USA[e]

Analysis of the genome sequence of the starlet sea anemone, *Nematostella vectensis*, reveals many genes whose products are phylogenetically closer to proteins encoded by bacteria or bacteriophages than to any metazoan homologs. One explanation for such sequence affinities could be that these genes have been horizontally transferred from bacteria to the *Nematostella* lineage. We show, however, that bacterium-like and phage-like genes sequenced by the *N. vectensis* genome project tend to cluster on separate scaffolds, which typically do not include eukaryotic genes and differ from the latter in their GC contents. Moreover, most of the bacterium-like genes in *N. vectensis* either lack introns or the introns annotated in such genes are false predictions that, when translated, often restore the missing portions of their predicted protein products. In a freshwater cnidarian, *Hydra*, for which a proteobacterial endosymbiont is known, these gene features have been used to delineate the DNA of that endosymbiont sampled by the genome sequencing project. We predict that a large fraction of bacterium-like genes identified in the *N. vectensis* genome similarly are drawn from the contemporary bacterial consorts of the starlet sea anemone. These uncharacterized bacteria associated with *N. vectensis* are a proteobacterium and a representative of the phylum *Bacteroidetes*, each represented in the database by an apparently random sample of informational and operational genes. A substantial portion of a putative bacteriophage genome was also detected, which would be especially unlikely to have been transferred to a eukaryote.

The starlet sea anemone, *Nematostella vectensis*, is a tiny, translucent invertebrate animal living in the estuarine salt marshes along both coasts of the United States, as well as the United Kingdom and Nova Scotia. *N. vectensis* belongs to the phylum *Cnidaria*, a primitive, nonbilaterian clade of metazoa, and is an increasingly popular model system for the study of the genetic control of metazoan body plan formation, the mechanisms of regeneration, and the evolution of development. A shotgun genome sequence of *N. vectensis* was obtained, assembled into long scaffolds, and annotated in 2007, and a remarkable resemblance of the gene repertoire in the species to the gene sets in more complex metazoa was emphasized, including a higher sequence similarity, a larger content of shared genes, and a higher degree of synteny between *N. vectensis* and vertebrates than between vertebrates and familiar invertebrate model organisms, such as the fruit fly and soil nematode (1). Detailed, case-by-case analysis of regulatory and signaling pathways shared by *N. vectensis* and higher animals has confirmed a premetazoan origin for many of these pathways, either in substantially complete form or in simpler versions that have been elaborated later in metazoan evolution by molecular "tinkering" (2, 3).

The other subset of *N. vectensis* genes, i.e., those that are neither metazoan inventions nor even eukaryote specific, has received relatively little attention. The ancient genes shared by a metazoan and prokaryotes are seen as the determinants of ancestral functions, vertically inherited from the prokaryotic ancestor of eukaryotes, as well as from the ancient alphaproteobacterium that must have given rise to modern-day mitochondria in the process of symbiogenesis (4). Such genes predate the origin of multicellular eukaryotes and are expected to be found, not only in bacteria and *N. vectensis*, but also in unicellular eukaryotes and different metazoa, allowing for gene losses in some of these lineages.

Recently, however, it has been noted that some genes in *N. vectensis* have orthologs in many bacteria but hardly any orthologs in other eukaryotes. For example, a trio of genes encoding putative cyclodipeptide synthases, the aminoacyl-tRNA synthase-like enzymes once thought to be restricted to bacteria, have been reported in *N. vectensis* (as well as one fungus and a lophotrochozoan invertebrate, but not in any other eukaryote), and one of the enzymes was expressed and shown to be enzymatically active (5).

In another study, Starcevic et al. (6) identified several genes in *N. vectensis* encoding putative enzymes of the shikimate pathway for biosynthesis of aromatic acids, which are not known to be produced by any metazoa. These sequences were too specifically related to the homologs from contemporary flavobacteria to attribute their origin to pre-eukaryotic or mitochondrial ancestry. Starcevic et al. noted that some of these genes contained introns (suggesting that the sea anemone may have acquired them from flavobacteria in times ancient enough to allow some intronization events), but other genes had no introns and possessed very high levels of sequence identity and codon usage similar to that of the bacterial homologs. This, as well as the additional gene sequences in the *N. vectensis* genome that had specific phylogenetic affinity to

Received 20 May 2013 **Accepted** 20 August 2013

Published ahead of print 30 August 2013

Address correspondence to Irena I. Artamonova, irenart@gmail.com, or Arcady R. Mushegian, mushegian2@gmail.com.

* Present address: Arcady R. Mushegian, Division of Molecular and Cellular Biosciences, National Science Foundation, Arlington, Virginia, USA.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/AEM.01635-13.

doi:10.1128/AEM.01635-13

flavobacteria or, in the case of homologs of bacterial 16S rRNA, to another bacterial lineage, pseudomonads, led Starcevic et al. to a hypothesis that the genome project of *N. vectensis* may have also sequenced portions of bacterial genomes present in the laboratory specimen of the sea anemone from which the genomic DNA library was prepared. Rather than dismissing these putative bacteria as biologically irrelevant sample contamination, Starcevic et al. suggested that they may represent novel, previously unknown consorts of the sea anemone, perhaps even its obligatory endosymbionts (6).

Our own curiosity about bacterium-like genes annotated as belonging to *N. vectensis* was piqued in the course of an unrelated project of sequence analysis of the global bacteriophage gene repertoire (7). We were particularly interested in genes with high phageness quotients, i.e., those genes that are often found in phages but almost never occur in bacterial genomes, except for the integrated prophages (7). When the all-inclusive sequence databases, such as the NT/NR databases at NCBI (8), are searched using these phage-specific protein sequences as queries in the BLAST family of algorithms (9), it is expected that the significant matches will originate from phage and prophage genomes, but not from their bacterial hosts, let alone eukaryotes. Surprisingly, some phage-specific genes also produced high-scoring matches to putative *N. vectensis* sequences. In agreement with Starcevic et al. (6), we could explain this in two ways: first, these genes may have been acquired by the *N. vectensis* genome in the evolutionary past by horizontal gene transfer from bacteria; second, there may be distinct bacterial and phage (or prophage) consorts in the laboratory lines of the starlet sea anemone. Obviously, either one or both of these phenomena could have contributed to the collection of *N. vectensis* genes of noneukaryotic origin, and it is of interest to know the relative contribution of each.

In this work, we provide several complementary lines of computational evidence that argue strongly for the existence of bacteria and viruses closely associated with *N. vectensis*. A substantial, perhaps nearly randomly sampled portion of this consortium's metagenome is already deposited in the databases, apparently having been misannotated as *N. vectensis* genes when in fact this gene complement should be studied further as evidence of the holobiont organization of the sea anemone.

MATERIALS AND METHODS

Genome sequencing data were downloaded from the Joint Genome Institute (JGI) home page (http://genome.jgi-psf.org/Nemve1). Nonglobular sequence segments were masked in the predicted *N. vectensis* proteins using the SEG program with settings optimized for the purpose (10, 11). The masked proteins were used as queries in searches by the BLASTP program (9) against the nonredundant protein subsection of the GenBank database using the E-value threshold (-e parameter) set to 10^{-6}. Protein taxonomic assignments were performed by comparison of the best matches in different taxa, as explained in more detail below.

The hypothesis of the overrepresentation of bacterium-like genes located in separate scaffolds was tested as follows. We randomly shuffled the classification labels of all genes from the *N. vectensis* genome project 1,000 times and calculated, for each such shuffling result, the number of genes with bacterial labels located in scaffolds that had no genes with eukaryotic labels. The number of such genes annotated in the real genome sequence data was then compared with the 0.01% upper quantile of the obtained distribution.

The GC content of the genomic scaffolds was calculated after masking the interspersed genomic repeats with the RepeatMasker program (http://www.repeatmasker.org).

Validation of the intron annotations in bacterium-like genes was done using three main criteria: (i) if the translation product of a bacterium-like, intron-containing gene could be mapped on a protein ortholog from the NR database along more than 90% of the query length without gaps longer than 3 amino acids, its predicted introns were considered to be true; (ii) if the predicted intron of a bacterium-like gene was flanked by two exons that mapped on the distal portions of the same protein sequence, such an intron was considered to be false, because in all such cases, the intron sequence translations restored the absent portion of the protein (see below); (iii) if the intron was flanked by two exons that mapped on different proteins in the same bacterial species, the intron was also considered false, as such predicted proteins are likely to be the artifacts of genome assembly or annotation; (iv) in cases where all database sequence matches were to protein repeats with moderate sequence similarity, the protein was removed from further examination.

To assess the phylogenetic positions of bacterium-like proteins more precisely, we selected all sequence matches obtained by the program BLASTP with E-values within 20 orders of magnitude from the best nonself match. For each taxonomic order that was found among the matches, two representatives were randomly chosen for phylogenetic tree construction (in rare cases, the sole available representative of the order had to be used). Trees were built based on the multiple protein sequence alignments produced by the ClustalW program (12) using the PhyML software (13) with 1,000 bootstrap replicates, with the parameter optimization mode employed at the default settings. Trees were visualized with the MEGA5 package (14). The shallowest tree clades that contained the query protein, had bootstrap support higher than 75%, and contained only a protein(s) from an organism(s) with a defined taxonomic position were considered to be phylogenetically informative tree neighborhoods of the query gene product.

The functions of the proteins encoded by *Pseudomonas mendocina* ymp and *Flavobacteria* bacterium BAL38, two genomes providing the maximal number of best database matches and closest phylogenetic neighbors for *N. vectensis* bacterium-like proteins, were assigned on the basis of the NCBI COG functional supergroups (15). We compared the Spearman's rank correlation coefficients between functional breakdowns for the whole proteome and the set of best matches for *N. vectensis* proteins.

To reannotate genes encoded by the bacteriophage scaffolds, we used the BLASTX and PSI-BLAST programs from the BLAST package (9), as well as Phage RAST (16).

RESULTS AND DISCUSSION

A considerable fraction of *N. vectensis* genes encode proteins of bacterial origin. The current annotation of the *N. vectensis* genome comprises 24,780 protein-coding genes. We masked the low-complexity sequence segments with the SEG program (11) and searched the NCBI protein NR database using the BLASTP program (9). More than one-fourth of the initial gene set (6,256 genes), when masked, had no nonself protein matches in NR. We assigned the rest of the gene products to the major domains of life (*Eukarya*, *Bacteria*, *Archaea*, or viruses) based on the consistent phylogenetic positions of the group of their highest-scoring database matches. A protein was ascribed to a particular domain if it met at least two of three surrogate criteria: (i) it had the best BLASTP match from this domain with an E-value of less than 10^{-6}, and any match from another domain had an E-value at least 2 orders of magnitude higher; (ii) the bit score for the best match in this domain was at least 50 bits higher than for any match in the other domains; (iii) within 20 orders of magnitude from the E-value of the best BLASTP match, the number of matches to proteins from the domain was at least two times larger than the number of such matches to proteins of any other domain. As a result, we provisionally classified 17,646 *N. vectensis* proteins. Manual

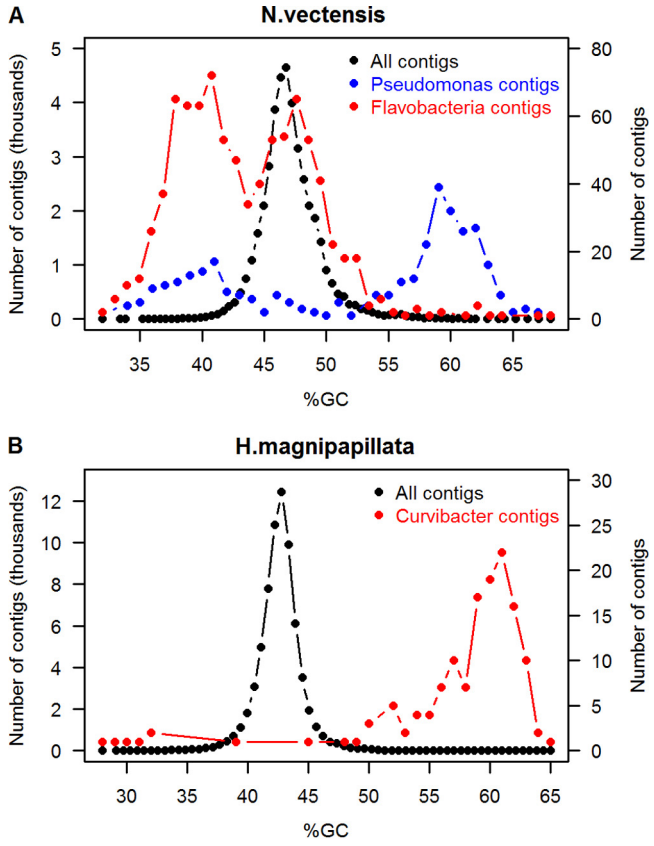

FIG 1 GC contents of *N. vectensis* (A) and *H. magnipapillata* (B) contigs. Only contigs from scaffolds with at least one annotated gene are included. In panel A, the curves corresponding to *Pseudomonas* and *Flavobacteriales* contigs are linked to the right vertical axis.

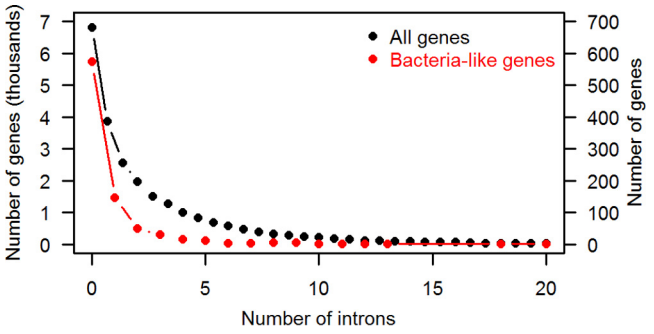

FIG 2 Distributions of intron numbers in the complete set and the bacterium-like subset of genes in the *N. vectensis* genome database. Only genes with less than 31 predicted introns are shown.

curation allowed us to classify 592 additional proteins. All told, 17,370 proteins were classified as eukaryotic, 850 as bacterial, 5 as archaeal, and 13 as viral (including bacteriophages), and 286 could not be classified with this protocol. The initial set of automatically assigned putative virus proteins consisted of seven bacteriophage proteins and six proteins from eukaryotic viruses. Case-by-case reanalysis of sequence similarities, including the verification of conserved sequence motifs, suggested that among six matches to eukaryotic viruses, only four were significant similarities, all of them to baculovirus glycosyltransferases (the closest database matches were GI:118197568, GI:68304238, and GI:215401455), whereas the others were spurious matches due to the presence of degenerate amino acid repeats not detected by the default BLAST composition-based statistics and the SEG filtering program (11). In contrast, all seven bacteriophage proteins had consistent similarity patterns, suggesting that they were closely related to gene products of podovirus 3 and two *Pseudomonas* phages, PaP2 and 119X.

Furthermore, 850 of the *N. vectensis* predicted proteins either had bacterial proteins as the closest match (typically much closer to the bacterial proteins than to eukaryotic homologs) or, in 42% of the cases, had significant matches only in bacteria but no matches to eukaryotic proteins below the E-value of 10^{-6}. The phylogenetically closest database matches of these bacterium-like sequences originated from several clades of bacteria, but nearly two-thirds of them belonged to one of two clades: proteobacteria (346 proteins, 151 of them from species of the genus *Pseudomonas*, i.e., *P. mendocina* or others) and bacteroidetes (356 proteins, 277 of them from the order *Flavobacteriales*). These proteins are listed in Table S1 in the supplemental material and are referred to as "bacterium-like" here.

The bacterium-like and eukaryotic proteins in *N. vectensis* reside on separate scaffolds. We examined the contents of the genomic scaffolds containing 850 bacterium-like proteins in *N. vectensis* and found that 670 of the proteins were encoded on the scaffolds containing no genes of eukaryotic provenance. This result is highly unexpected under the hypothesis of random distribution of bacterium-like genes among all protein-coding scaffolds in *N. vectensis* ($P < 0.0001$). We tentatively classified the contigs and scaffolds by the predominant phylogenetic affinity of the proteins encoded by these genome fragments. A contig or scaffold was classified as eukaryotic if it encoded one or more eukaryotic proteins, as bacterium-like if it encoded at least one bacterium-like protein and no eukaryotic proteins, as archaeon-like if it encoded an archaeal protein(s) and no eukaryotic proteins, and as viral if it encoded a viral protein(s) and no proteins of another origin. Using these criteria, we classified 3,225 genomic scaffolds of *N. vectensis* that included at least one gene model with a phylogenetically assigned product. Among these scaffolds, 2,705 were classified as eukaryotic, 517 as bacterial, and 3 as phage or prophage; there were no archaeal scaffolds.

Analysis of the nucleotide compositions of the scaffolds showed that eukaryotic and bacterium-like scaffolds in *N. vectensis* have distinct distributions of the GC contents, which are shown in Fig. 1A. It is notable that the distributions of the GC contents of bacteroidetes-like and proteobacterium-like scaffolds track closely with the compositional properties of the respective bacterial genomes and are very different from the eukaryotic scaffolds, suggesting a lack of nucleotide sequence amelioration (17) of bacterium-like genes deposited in *N. vectensis* genome databases.

Many introns in bacterium-like genes from the *N. vectensis* genome project are falsely predicted. The majority of genes in the *N. vectensis* genome (72%) are predicted to contain at least one intron. Interestingly, however, there are pronounced differences in the statistics of introns in eukaryotic and bacterium-like scaffolds. Indeed, almost 75% of the genes in eukaryotic scaffolds contain introns—more than 6 introns per gene on average—while only 23% of the genes in bacterium-like scaffolds are predicted to contain introns—only 1.6 introns per gene on average (Fig. 2).

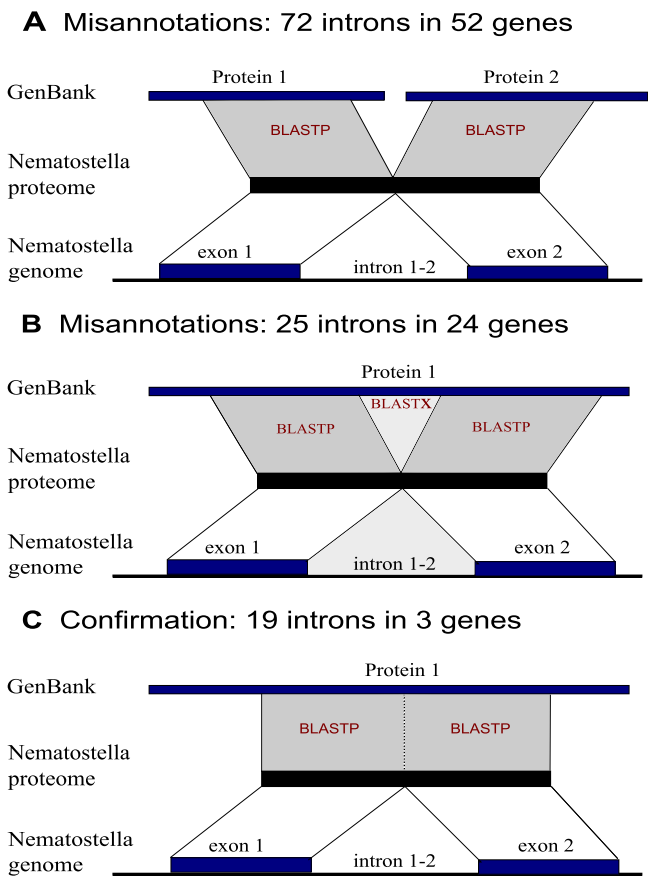

FIG 3 Intron validation for genes located in bacterium-like scaffolds. The numbers of introns and genes in each set are indicated. In addition, 14 genes were misclassified during this analysis, as all their BLASTP matches originated from protein repeats or similar short domains.

Closer examination indicated that among 251 introns annotated for bacterium-like genes, in 147 cases the protein level match could be extended into the putative "intron." A case-by-case examination demonstrated that 113 "introns" in 97 genes contained contig borders (NNN tracks) within them. All of these introns seem to be falsely predicted, because as the NNN sequence cannot be incorporated into CDS regions, the motifs similar to splicing sites were found in the flanking regions of each NNN track.

For the remaining 138 completely sequenced introns from 93 genes, a more detailed analysis was done. They were inspected by comparing their translation products with the proteins that matched their flanking exons, as illustrated in Fig. 3. All these bacterium-like, intron-containing genes would be "smoking guns" for eukaryotic acquisition and "domestication" if the introns in them were genuine. Far from this being the case, however, 55 "introns" from 46 genes could be translated into a product with significant similarity to bacterial proteins. In 38 of these cases, the putative product was an extension of the product encoded by the adjoining "exons." Only 19 introns in three genes were confirmed by our analysis to be true introns; however, in all these cases, we have further discovered that the genes had been misclassified as bacterial, as they have best matches that belong to bacteria but that are only marginally better than multiple matches from eukaryotes.

Phylogenetic confirmation of bacterium-like proteins. We validated our protein taxonomic assignments further by collecting the database homologs of each query protein, constructing more accurate multiple alignments than those produced by BLASTP database searches, and inferring phylogenetic trees using the maximum-likelihood method (see Materials and Methods). For 54 bacterium-like proteins, the set of database homologs within the vast sequence similarity range consisted of only one (bacterial) protein. Twenty-three such matches originated from *Flavobacteriales* and two from pseudomonads. Phylogenetic trees could be constructed for the other 796 proteins, and in 636 cases, the bacterium-like proteins from the *N. vectensis* genome project were clustered with other bacterial proteins in the tree. In particular, 234 proteins were clustered with flavobacterial and 121 with pseudomonad homologs. Only 17 bacterium-like proteins were found to cluster with eukaryotic proteins from the database. For 140 proteins, phylogenetic evidence was inconclusive, as the *N. vectensis* proteins belonged to clades that included a mix of prokaryotic and eukaryotic proteins with unclear branching order.

Functional classes of bacterium-like proteins in *N. vectensis* are apparently randomly sampled from the bacterial genomes. The majority of bacterium-like genes from the *N. vectensis* genome project are phylogenetically closer to *Bacteroidetes* or *Proteobacteria* homologs than to any eukaryotic homolog. More specifically, almost all of these proteins have their best BLASTP matches in *Flavobacteriales* or *Pseudomonas* at high levels of amino acid identity (80 to 85% and 85 to 95% median identity, respectively) to their nearest database homologs from these clades.

We analyzed the functions of these proteins by examining the functional annotation of their close matches and comparing them with the partitioning of the complete bacterial proteomes into NCBI COG supercategories (15). Spearman's rank correlation statistics suggest that distributions of functions in *N. vectensis* bacterium-like protein sets resemble the functional partitioning of the complete proteomes of their closest bacterial relatives much more than the partitioning of the *N. vectensis* eukaryotic protein set (the respective correlation coefficients are 0.88 versus 0.75 in the case of *P. mendocina* ymp and 0.86 versus 0.60 for *Flavobacteria* bacterium BAL38). This difference indicates that the sets of bacterium-like proteins reported by the *N. vectensis* genome project are more likely to have been randomly sampled from the corresponding bacterial proteomes than recruited into *N. vectensis* because of their specific molecular functions.

Many of the bacterium-like proteins annotated by the *N. vectensis* genome project have functions that are relevant to the biology of bacteria but have never been reported in eukaryotes. Examples are proteins with high similarity to Cas1 (one of the CRISPR-associated proteins participating in a prokaryote-specific immunity system), bacterial transcription regulators and bacterial-type protein kinases, the pilin glycosylation protein PglD, OmpA/MotB domain proteins that operate in the outer membrane of Gram-negative bacteria, and the aforementioned cyclopeptide biosynthesis enzymes. Even if the genes encoding these proteins were once horizontally transferred to the *N. vectensis* genome, it would be highly unusual for a eukaryote to maintain all of them, in an essentially unmodified form, to perform a set of functions that are without precedent in metazoa. Sequencing of the extant bacterial genomes by the scientists working on the *N. vectensis* genome project, rather than acquisition of genes by the *Nematostella* lineage, seems to the most plausible explanation for such observations.

Particularly unusual were seven putative *N. vectensis* proteins

confidently classified as close relatives of bacteriophage proteins on the basis of their highly significant sequence similarity to homologs from *Pseudomonas* phage PaP2, *Pseudomonas* phage 119X, and EBPR podovirus 3. All these phages belong to the family *Podoviridae*, and two of them are known to infect *Pseudomonas* species. We reannotated three genomic scaffolds on which these proteins were encoded and found a total of 32 open reading frames (ORFs), mostly located close to each other on the same strand in each scaffold, and no bacterium-like or eukaryotic proteins. Twenty-four of these ORF products have homologs encoded by other phage genomes, and for eight of them, we were able to predict general functions (see Table S2 in the supplemental material).

The bacterium-like genes revealed by the *N. vectensis* genome project partition from the host genes similarly to the genes of the known bacterial endosymbiont of *Hydra magnipapillata*. In our analysis of putative bacterial genes that have been sequenced in the course of the *N. vectensis* genome project, we have obtained several kinds of evidence that, in our opinion, are most compatible with the idea that the sea anemone genomic library from the CH2 × CH6 laboratory strain contained sequences from at least two distinct bacteria. These bacterial species are likely to be closely associated with the *N. vectensis* individuals that were used as the DNA source, and also perhaps with the body of sea anemones in the wild. No such bacterial inhabitants of the sea anemone have been reported in the literature, let alone cultivated. Interestingly, however, many properties of bacterial scaffolds deposited by the *N. vectensis* genome project are similar to the set of scaffolds with bacterial genes obtained more recently by a genome project of another cnidarian, *Hydra magnipapillata*, where the bacterial endosymbiont has in fact been described (18). Our classification protocol applied to the *H. magnipapillata* genome revealed that among the 17,385 gene products, there were 401 proteins classified as bacterial. The majority of these proteins have already been recognized as belonging to the bona fide proteobacterial endosymbiont of *H. magnipapillata* (19). Interestingly, our protocol also detected two proteins (GI:449662904 and GI:449662902) highly similar to protein products of giant DNA viruses from the megavirus and moumouvirus groups, suggesting that the *Hydra* microbiome may include additional life forms.

Similar to what we observed with the *N. vectensis* bacterium-like genes, the genes of the bacterial endosymbiont of *H. magnipapillata* tend to be intronless open reading frames, which are located in separate genomic scaffolds characterized by distinct GC contents. The *Hydra* genome project appears to have few, if any, intron prediction artifacts in these genes, perhaps because the presence of a bacterial symbiont in this biological specimen was known in advance and corresponding genes could be verified with bacterium-specific models.

We extended earlier observations (6) to show that the *N. vectensis* genome sequence submitted to the sequence databases is very likely to contain many genes of bacterial and bacteriophage origin. They are located in distinct scaffolds, which can be separated from the DNA of the target organism, *N. vectensis*, by several criteria, including the nucleotide composition and the provenance of neighboring genes; the spurious character of introns annotated in these genes; and a rich and apparently randomly drawn repertoire of predicted protein functions, which include several molecular roles considered to be prokaryote specific. All these lines of evidence suggest that of the two possible explanations for the presence of these genes in the data, i.e., (i) domestication by an invertebrate of horizontally transferred bacterial genes in the evolutionary past and (ii) the occurrence of bacteria in the *N. vectensis* planula used for DNA isolation, the latter seems to be more plausible.

We propose that, far from these genes being biologically irrelevant sample contamination, their presence in the genome database indicates that, similar to the cases of other marine *Anthozoa* (20–22) and the completely sequenced freshwater hydrozoan *H. magnipapillata*, the body of the starlet sea anemone is in fact a holobiont, i.e., a consortium of a metazoan animal and bacteria closely associated with it throughout most or all of the host life cycle. As far as we know, no firm microbiological evidence for such consorts of *N. vectensis* has been published.

Two distinct bacterial genera are clear leaders in high sequence similarity of these bacterium-like proteins. They are *Pseudomonas* and *Flavobacterium*, separated by many hundreds of millions of years of bacterial evolution. Interestingly, the alternative sequence assembly available in the StellaBase database (http://stellabase.org/; 23) contains 16S RNA from exactly these two genera (6).

On a more general note, the idea of horizontal gene transfer from bacterial endosymbionts to their eukaryotic hosts, which sounded heretical a decade ago, is now widely accepted. It stands to reason that at least some bacterial genes may be "domesticated" by the host, perhaps because of the selective advantage of the expanded metabolic capability that these genes confer (5, 24–26). However, in other situations, such as the one described here, the presence of microbial genes in a draft eukaryotic genome assembly may provide, "at the point of the investigator's pen," the clues to the host-microbe associations that take place today but have evaded detection by more traditional microbiological methods.

ACKNOWLEDGMENTS

We thank Evgeniy N. Gordienko and Anna A. Gogleva for technical support.

The views expressed in this article are ours and do not necessarily represent the views of the NSF or the U.S. government.

REFERENCES

1. **Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS.** 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science **317:** 86–94.
2. **Ryan JF, Mazza ME, Pang K, Matus DQ, Baxevanis AD, Martindale MQ, Finnerty JR.** 2007. Pre-bilaterian origins of the Hox cluster and the Hox code: evidence from the sea anemone, *Nematostella vectensis*. PLoS One **2:**e153. doi:10.1371/journal.pone.0000153.
3. **Adamska M, Larroux C, Adamski M, Green K, Lovas E, Koop D, Richards GS, Zwafink C, Degnan BM.** 2010. Structure and expression of conserved Wnt pathway components in the demosponge *Amphimedon queenslandica*. Evol. Dev. **12:**494–518.
4. **Andam CP, Williams D, Gogarten JP.** 2010. Natural taxonomy in light of horizontal gene transfer. Biol. Philos. **25:**589–602.
5. **Seguin J, Moutiez M, Li Y, Belin P, Lecoq A, Fonvielle M, Charbonnier JB, Pernodet JL, Gondry M.** 2011. Nonribosomal peptide synthesis in animals: the cyclodipeptide synthase of *Nematostella*. Chem. Biol. **18:** 1362–1368.
6. **Starcevic A, Akthar S, Dunlap WC, Shick JM, Hranueli D, Cullum J, Long PF.** 2008. Enzymes of the shikimic acid pathway encoded in the genome of a basal metazoan, *Nematostella vectensis*, have microbial origins. Proc. Natl. Acad. Sci. U. S. A. **105:**2533–2537.
7. **Kristensen DM, Cai X, Mushegian A.** 2011. Evolutionarily conserved orthologous families in phages are relatively rare in their prokaryotic hosts. J. Bacteriol. **193:**1806–1814.
8. **Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW.** 2012. GenBank. Nucleic Acids Res. **40:**D48–D53.

9. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
10. **Wootton JC.** 1994. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput. Chem. **18:** 269–285.
11. **Wootton JC, Federhen S.** 1996. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. **266:**554–571.
12. **Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG.** 2007. Clustal W and Clustal X version 2.0. Bioinformatics **23:**2947–2948.
13. **Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O.** 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. **59:**307–321.
14. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S.** 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. **28:**2731–2739.
15. **Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA.** 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:**41.
16. **Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O.** 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics **9:**75.
17. **Lawrence JG, Ochman H.** 1997. Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. **44:**383–397.
18. **Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, Rattei T, Balasubramanian PG, Borman J, Busam D, Disbennett K, Pfannkoch C, Sumin N, Sutton GG, Viswanathan LD, Walenz B, Goodstein DM, Hellsten U, Kawashima T, Prochnik SE, Putnam NH, Shu S, Blumberg B, Dana CE, Gee L, Kibler DF, Law L, Lindgens D, Martinez DE, Peng J, Wigge PA, Bertulat B, Guder C, Nakamura Y, Ozbek S, Watanabe H, Khalturin K, Hemmrich G, Franke A, Augustin R, Fraune S, Hayakawa E, Hayakawa S, Hirose M, Hwang JS, Ikeo K, Nishimiya-Fujisawa C, Ogura A, Takahashi T, Steinmetz PR, Zhang X, Aufschnaiter R, Eder MK, Gorny AK, Salvenmoser W, Heimberg AM, Wheeler BM, Peterson KJ, Bottger A, Tischler P, Wolf A, Gojobori T, Remington KA, Strausberg RL, Venter JC, Technau U, Hobmayer B, Bosch TC, Holstein TW, Fujisawa T, Bode HR, David CN, Rokhsar DS, Steele RE.** 2010. The dynamic genome of *Hydra*. Nature **464:**592–596.
19. **Fraune S, Bosch TC.** 2007. Long-term maintenance of species-specific bacterial microbiota in the basal metazoan *Hydra*. Proc. Natl. Acad. Sci. U. S. A. **104:**13146–13151.
20. **Williams GP, Babu S, Ravikumar S, Kathiresan K, Prathap SA, Chinnapparaj S, Marian MP, Alikhan SL.** 2007. Antimicrobial activity of tissue and associated bacteria from benthic sea anemone *Stichodactyla haddoni* against microbial pathogens. J. Environ. Biol. **28:**789–793.
21. **Xiao H, Chen Y, Liu Z, Huang K, Li W, Cui X, Zhang L, Yi L.** 2009. Phylogenetic diversity of cultivable bacteria associated with a sea anemone from coast of the Naozhou island in Zhanjiang, China. Wei Sheng Wu Xue Bao **49:**246–250.
22. **Du Z, Zhang W, Xia H, Lü G, Chen G.** 2010. Isolation and diversity analysis of heterotrophic bacteria associated with sea anemones. Acta Oceanol. Sin. **29:**62–69.
23. **Sullivan JC, Reitzel AM, Finnerty JR.** 2008. Upgrades to StellaBase facilitate medical and genetic studies on the starlet sea anemone, *Nematostella vectensis*. Nucleic Acids Res. **36:**D607–D611.
24. **Hall C, Brachat S, Dietrich FS.** 2005. Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. Eukaryot. Cell **4:**1102–1115.
25. **Nikoh N, Nakabachi A.** 2009. Aphids acquired symbiotic genes via lateral gene transfer. BMC Biol. **7:**12.
26. **Ros VI, Hurst GD.** 2009. Lateral gene transfer between prokaryotes and multicellular eukaryotes: ongoing and significant? BMC Biol. **7:**20.